# Computational methods for the analysis of bacterial gene regulation

Brouwer, Rutger Wubbe Willem

# *Chapter 7*

## *General discussion*

## *Summary*

The elucidation of the complete genome sequences of both *L. lactis* subsp. *lactis* IL1403 [6] and *L. lactis* subsp. *cremoris* MG1363 [8] allowed the development and use of state-of-the-art DNA microarrays [21,101,103,108,144], proteomics (2D gel electrophoresis and mass spectrometry) [94,104], and genomics tools [96,145]. For both *L. lactis* genomes, curated metabolic models have been made [9,96]. Even though these advancements have greatly contributed to understanding the *L. lactis* physiology, its gene content and regulation, many aspects of the biology of *L. lactis* still remain to be uncovered. For example, little is known on the roles of the putative transcriptional regulators encoded in the *L. lactis* MG1363 genome [112].

In this thesis, computational methods were developed and used to expand our knowledge on the regulation of gene transcription in bacteria and specifically *L. lactis* MG1363. To this end, detailed studies were conducted into operon prediction methods that predict the basic transcriptional units in the bacterial cell (Chapter 2 and [35]). Further analysis of the operon predictions revealed the best genomic properties on which to base these predictions (Fig. 1) as well as the importance of a suitable algorithm to integrate this knowledge (Fig. 1; Chapter 3). A web-tool for genome-centric data visualization, MINOMICS, is introduced in Chapter 4 [146]. In Chapter 5, gene expression in *L. lactis* MG1363 grown in rich media during batch fermentation was investigated through a high-density DNA microarray time-course experiment. In this time-course, gene regulation events were observed for key biological systems including amino-acid and nucleotide metabolism. Analysis of this time-course data with advanced bioinformatics and graph analysis tools (Fig. 1) allowed generating a genetic network for *L. lactis* MG1363 that is presented in Chapter 6. In this network, co-expressed genes in the *L. lactis* MG1363 time-course were clustered using a clique-based graph approach. Quasi-cliques in this network are analogous to regulons. The network allows extending existing regulons as well as postulating new regulon structures for *L. lactis* MG1363.

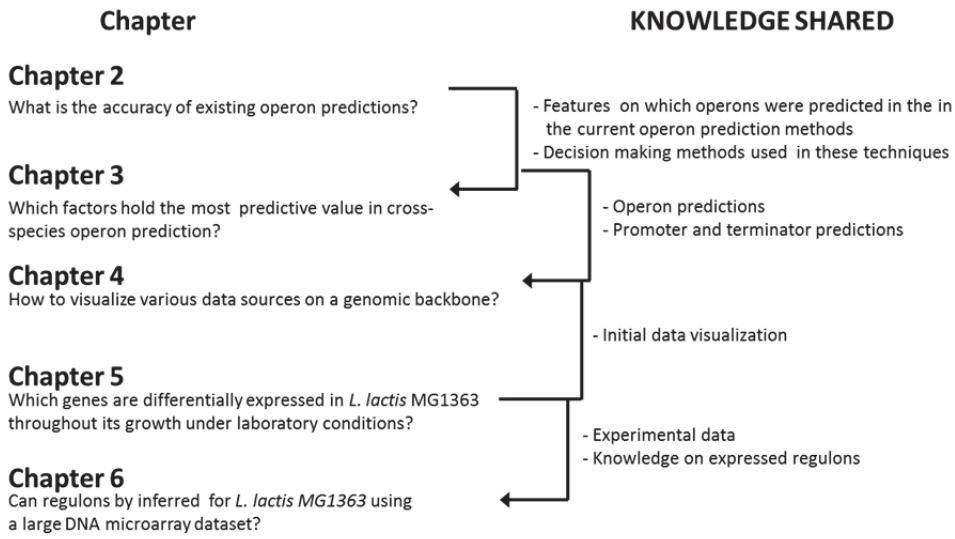


Fig. 1 Diagram on how knowledge is generated and shared between the chapters of this thesis
Left-hand side: chapters and their leading research question. Right-hand side: most important connections between these chapters are shown.

## *Transcriptional organization in bacteria*

The first operon predictions appeared in literature in 2000, shortly after the first whole genome sequences became available [52]. Traditionally, this field of bioinformatics has attracted many computer scientists, since the problem of predicting operons is seemingly ideal for machine learning. The main question to answer is whether two neighboring genes are part of the same transcript [55]. In machine learning terms, this question is a simple two class prediction problem where the classes (transcriptional unit or not in operon) can be predicted using one or more properties of the considered gene pair (features). Any property that can be determined for two adjacent genes can potentially be used as feature and may contribute to predicting whether two adjacent genes are in an operon. (for a review see [35]). For operon prediction, features such as gene co-inheritance and intergenic distance were successfully used to predict operon membership (Chapter 2). These features are relatively easy to obtain from the genome sequence and annotation for any given bacterial genome. In addition to features, machine learning based predictions require

training data. For operon predictions, this training data consists of experimentally verified operons which are available for multiple organisms including *E. coli* [80] and *B. subtilis* [19]. Due to the above-mentioned reasons, the operon prediction field has been very successful in the past years and will be relevant as long as (new) bacterial genomes become available. Operon predictions have been described that utilized many different features that were combined using numerous learning techniques ranging from logistic classifiers to neural networks (Chapter 2).

Even though numerous operon prediction methods have been developed (for a review see Chapter 2 [35]), no method has reached 100 % accuracy, not even for the extensively characterized model organisms *E. coli* K12 or *B. subtilis* 168. The top predictor for gene-pairs to be part of the same transcriptional unit is the intergenic distance (Chapters 2 and 3) with the following rationale: at shorter intergenic distances, there is little space to encode transcriptional control signals, such as transcriptional terminators and promoters and transcription factor binding sites. For this reason, genes that are transcribed in different lie generally over 50 base pairs apart on the genome. Operon prediction methods that do not take into account the intergenic distance [16,44,57,69] have thus far not been very successful even when the operon predictions were based on the inference of many other properties that could be used to predict membership of an operon (*i.e.* presence of promoters in the upstream region, terminators, and transcription factor binding sites) (Chapter 2). Therefore, we hypothesize that the elements controlling gene transcription in bacteria can for a small part not (yet) be accurately determined entirely from the genome sequence. This is even true for the widely studied model organisms *E. coli* and *B. subtilis.*

The operon prediction problem is complicated by genes that are co-transcribed only under specific conditions [11]. One explanation of conditional operons is the occurrence of multiple promoters upstream and in the operon. These promoters are active under (slightly) different conditions resulting in different transcript and thus conditional operons. An alternative explanation would be the presence of conditional dependent transcriptional terminators [11,65,75]. These could operate by selectively recruiting the Rho complex to transcript. Only recently, RNA sequencing (RNA-seq) has become available. With some RNA-seq methods, cDNA transcripts are completely covered allowing the start and ends of the transcripts to be determined. With these techniques conditional operon structures can be investigated genome-wide [147]. As only a few conditional operons have been described, a dataset of sufficient size to predict conditional dependence of operons

is currently lacking. Therefore, conditional dependent operon structures have not explicitly been taken into account by any of the described operon predictions. In this thesis, both the quality of the available operon predictions were determined (Chapter 2) as well the predictive value of the used features (Chapter 3).

Similar to operon prediction, genetic network reconstruction could be presented as a relatively simple two-class problem, where the question is whether two genes are regulated by the same transcriptional regulator or not [129]. However, this problem is much harder to solve as there is no specific feature described that is particularly information rich [129]. Most genetic networks are determined from gene-to-gene correlations in large gene expression datasets supplemented with additional experimental information and/or transcriptional motif predictions [129]. The genes of each regulon should be differentially co-expressed to generate the complete genetic network from such data [129]. For most organisms, it is impractical or even impossible to obtain such a dataset as the conditions under which specific regulators operate on their target genes are not known or cannot be reproduced under laboratory conditions.

Transcription factor binding motif predictions can supplement gene expression based networks [129], but may not predict each motif effectively. Genome-wide chromatin immuno-precipitation (ChIP) datasets are a better resource to associate genes to their transcriptional regulators. In ChIP experiments, the genomic sites are identified to which a DNA-associated protein binds. In this procedure, a cell culture is treated with a reversible cross-linking agent which fixates proteins to the genomic DNA. The cross-linked material is then fragmented and purified. Using an antibody specific to the target protein, the protein-DNA complexes with this protein can be enriched. After reversing the cross-links, the resulting DNA can either by hybridized to DNA microarrays or sequenced yielding the genomic sites to which the protein was associated [131,148]. However, performing ChIP experiments on a large scale is in most cases not economically feasible.

In this thesis, a correlation network for *L. lactis* MG1363 was determined based on a detailed gene expression time-course dataset that was supplemented with a MEME transcription factor binding motif prediction (see Chapter 6). Throughout growth, the gene expression of many transcriptional regulators is changed presumably causing differential expression of many regulons (Chapter 5). This network was based on genes and operons that were correlated in expression. In this networks groups of co-expressed genes and operons were determined that are analogous to regulons. The groups in the network did not exactly match the regulons described in literature. These differences

can either be caused by genes that were erroneously reported to be part of regulons or by multiple transcriptional regulators influencing the transcription of these gene groups. Individual cases of multiple transcriptional regulators affecting the expression of genes have been widely reported also in *L. lactis* [101–103]. By accurately determining co-expressed genes, known regulons may be refined; if gene A is in a regulon and its expression is highly correlated to that of gene B, it is likely that gene B is also in the same regulon. Such a connection should be followed up with further bioinformatics evidence, such as shared DNA binding motifs, or experimental follow up studies.

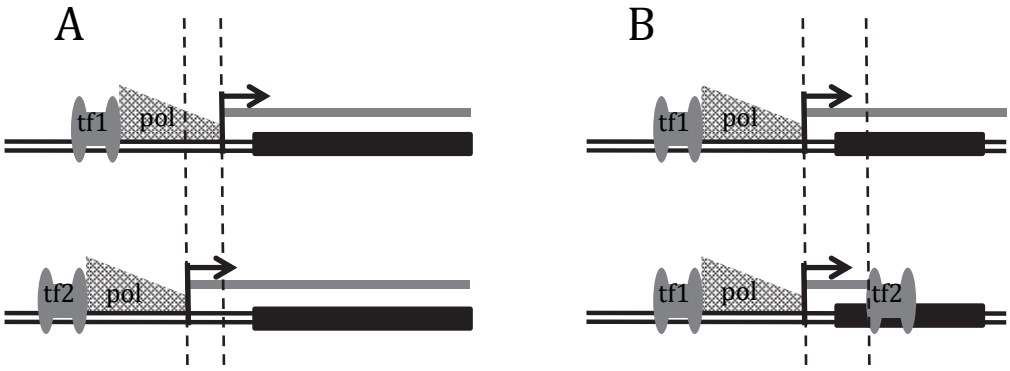


Fig. 2 Regulation events distinguishable by RNA sequencing Transcription factors (tf1 and tf2) bind to the promoter and recruit the RNA polymerase complex (pol). This complex starts transcription at the transcription start site (black arrow) and synthesizes the mRNA transcript (gray line). In situation A, 2 transcription factors bind at different positions in the genome resulting in 2 distinct transcription start sites. These start sites can be discerned from the 5' end of the transcript sequence (dotted line). In situation B, transcription factor 2 represses gene expression resulting in partially synthesized transcripts (dotted line). This transcript could potentially be picked up using RNA-seq.

Ideally, the combined contributions of different transcriptional regulators to the expression of a given gene should be quantified with a single experimental technique. However, such quantifications are difficult to perform with DNA microarray data as the probes of most DNA microarrays specify a few locations of a transcript. With RNA sequencing technology (RNA-seq) [147,149] specific transcripts can be distinguished. RNA-seq methods allow cDNA fragments to be generated

over the entire length of the transcript. The exact coverage and sequence of these fragments should allow for different regulation and RNA processing events to be discerned (example in Fig. 2)

Using RNA-seq data in combination with specific experimental designs, such as time-courses or designs in which many experimental parameters are varied could allow discerning regulatory interactions and could therefore be a large benefit to genetic network reconstruction. The experimental design should be aimed to maximize differential expression in response to variations in the medium and environment. The common reference in this experiment should be a chemically defined medium in which the organism can grow at near optimal speed. Relatively minor variations in the experimental parameters, such as nutrient concentration, pH and temperature, could then be used to trigger transcriptional responses. There are two practical challenges with such experimental designs. The first is to ensure that growth speed is not greatly affected because in that case the intended and more local transcriptional response cannot be discerned from a more global response caused by retarded growth. Second is that many parameters should be perturbed in order to reconstruct a clearly defined genetic network that covers a large portion of the regulons of the organism. The success of such a study will lie in no small part to balancing the costs to the potential (scientific) gains. When considering too many parameters, might not be cost-effective, but considering too few will result in insufficient resolution. .

By basing genetic networks on existing datasets, such as the *L. lactis* time-course experiment (Chapters 5 and 6), costs can be greatly reduced and many regulons can still be inferred. One issue with using these datasets is too integrate data from different platforms that have become more accurate over time. Sequencing based techniques are still costly and only a few methods are available for preparing sequence libraries from prokaryotic RNA [149–151]. Data analysis techniques for RNA-seq data are still evolving therefore requiring specialists to obtain the full benefit of the data [152,153]. The normalization methods and down-stream analysis techniques for DNA microarray are well established and understood. RNA-seq requires different normalization and analysis techniques from DNA microarray data as RNA-seq is count based and requires different statistical models. These models are currently being developed and refined [152–154]. However, RNA-seq offers a more complete picture of the RNA and allows identification of different transcription start sites as well as RNA decay. We foresee that, due to these analytical issues, DNA microarrays will remain the standard technique for measuring gene expression in prokaryotes in many

research groups for the near future. For more specialized research questions, RNA-seq will be the method of choice.

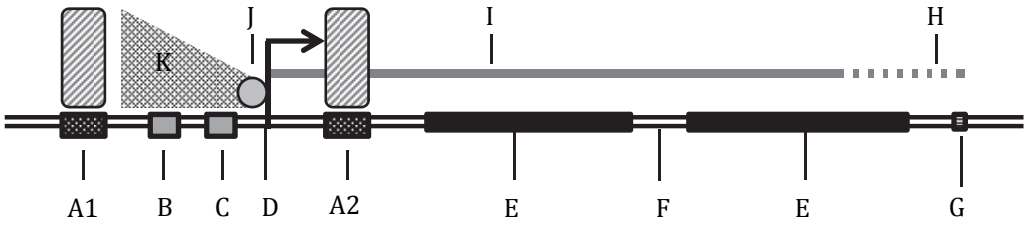


Fig. 3 Factors influencing gene-expression in bacteria
In this schematic overview, some of the factors influencing bacterial gene expression are listed. A) Transcription factor binding site, B) -35 sequence, C) -10 sequence, D) Transcription start site, E) Protein coding gene sequence, F) inter-genic region (not translated), G) Transcriptional terminator site, H) mRNA degradation, I) mRNA transcript J) sigma factor, K) RNA polymerase. The transcription factor at A1 may be an enhancer or a repressor. The transcription factor at A2 is a repressor working via the roadblock mechanism. Transcriptional termination (G) may be protein dependent and thus could be target for regulation.

## *Concluding remarks and future prospects*

The main subject of this thesis is the study of transcription of genes in bacteria in general and *L. lactis* MG1363 in particular. By using DNA microarrays, gene expression in these organisms can now be determined on a genome-wide scale yielding valuable insights in the underlying regulatory processes. Through the advent of next-generation sequencing, new techniques have been developed to study various other genetic and epigenetic aspects in eukaryotes and prokaryotes. Most of these techniques are not organism specific. These new techniques will significantly improve our understanding of prokaryotic gene regulation and epigenetics in the years to come. For example, chromatin immunoprecipitation sequencing (ChIP-seq) allows determination of the genome association of specific factors (Fig. 3: A1, A2, protein dependent G [65] and J) on a genome-wide scale [131]. By comparing the binding patterns of an activated and a non-activated transcriptional regulator, direct evidence for gene regulation is

generated, enabling inference of transcriptional control and inference of regulons. Massive parallel sequencing based techniques can also be used to determine DNA methylation patterns across the genome [155,156]. Methylation patterns have been shown to influence gene expression in eukaryotes and may also have effects in prokaryotes although these also employ other ways of epigenetic inheritance [157,158]. It would be interesting to see the effect of methylation on regulatory elements in prokaryotes (Fig. 3). Other techniques that explore chromosome structure, such as chromosome conformation capture [159], may not seem directly relevant to prokaryotic genetics since the structure of prokaryotic DNA is thought to have little impact on gene regulation and expression. However, these techniques may provide surprising results as the mechanisms behind DNA organization has only been recently been described in eukaryotes. To our knowledge these mechanisms have not been researched in bacteria. The field of prokaryotic genetics is fully benefiting from an influx of new techniques and methodologies that will greatly enhance our understanding of the prokaryotic genome and the regulation of its genes. Bioinformatics will surely be a key element in analyzing, assembling, comparing, and interpreting these new and exciting datasets.

# *References*


1. Madigan, M. T., Martinko, J. M., Dunlap, P. V & Clark, D. P. *Brock Biology of Microorganisms*. *Cell* **2**, 1168 (Pearson/Benjamin Cummings: 2009).
2. Woese, C., Dugre, D., Saxinger, W. & Dugre, S. The molecular basis for the genetic code. *Proceedings of the National Academy of Sciences of the United States of America* **55**, 966 (1966).
3. Blattner, F. R. The Complete Genome Sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
4. Kunst, F. & Devine, K. Sequencing project of *Bacillus subtilis* genome. *Research in Microbiology* **142**, 905–912 (1993).
5. Gasson, M. J. Genetic transfer systems in lactic acid bacteria. *Antonie van Leeuwenhoek* **49**, 275–82 (1983).
6. Bolotin, a *et al.* The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome research* **11**, 731–53 (2001).
7. Ventura, M. *et al.* Comparative analyses of prophage-like elements present in two *Lactococcus lactis* strains. *Applied and environmental microbiology* **73**, 7771–80 (2007).
8. Wegmann, U. *et al.* Complete genome sequence of the prototype lactic acid bacterium *Lactococcus lactis* subsp. *cremoris* MG1363. *Journal of bacteriology* **189**, 3256–70 (2007).
9. Siezen, R. J. *et al.* Complete genome sequence of *Lactococcus lactis* subsp. *lactis* KF147, a plant-associated lactic acid bacterium. *Journal of bacteriology* **192**, 2649–50 (2010).
10. Van Hijum, S. A. F. T., Medema, M. H. & Kuipers, O. P. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiology and molecular biology reviews MMBR* **73**, 481–509, Table of Contents (2009).
11. Okuda, S. *et al.* Characterization of relationships between transcriptional units and operon structures in *Bacillus subtilis* and *Escherichia coli*. *BMC genomics* **8**, 48 (2007).
12. Price, M. N., Arkin, A. P. & Alm, E. J. The life-cycle of operons. *PLoS genetics* **2**, e96 (2006).
13. Roback, P. *et al.* A predicted operon map for *Mycobacterium tuberculosis*. *Nucleic acids research* **35**, 5085–95 (2007).
14. Laing, E., Sidhu, K. & Hubbard, S. J. Predicted transcription factor binding sites as predictors of operons in *Escherichia coli* and *Streptomyces coelicolor*. *BMC genomics* **9**, 79 (2008).
15. Price, M. N., Arkin, A. P. & Alm, E. J. OpWise: operons aid the identification of differentially expressed genes in bacterial microarray experiments. *BMC bioinformatics* **7**, 19 (2006).
16. Zheng, Y., Szustakowski, J., Fortnow, L., Roberts, R. & Kasif, S. Computational identification of operons in microbial genomes. *Genome research* **12**, 1221–1230 (2002).

17. Redon, E., Loubière, P. & Cocaign-Bousquet, M. Role of mRNA stability during genome-wide adaptation of *Lactococcus lactis* to carbon starvation. *The Journal of biological chemistry* **280**, 36380–5 (2005).
18. Gama-Castro, S. *et al.* RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic acids research* **39**, D98–105 (2011).
19. Sierro, N., Makita, Y., De Hoon, M. & Nakai, K. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic acids research* **36**, D93–6 (2008).
20. Demeter, J. *et al.* The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic acids research* **35**, D766–70 (2007).
21. Even, S., Lindley, N. D. & Cocaign-Bousquet, M. Molecular physiology of sugar catabolism in *Lactococcus lactis* IL1403. *Journal of bacteriology* **183**, 3817–24 (2001).
22. Yi, H. *et al.* Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic acids research* **39**, e140 (2011).
23. Yang, Y. H., Buckley, M. J. & Speed, T. P. Analysis of cDNA microarray images. *Briefings in bioinformatics* **2**, 341–9 (2001).
24. Yang, Y. H. & Speed, T. Design issues for cDNA microarray experiments. *Nature reviews. Genetics* **3**, 579–88 (2002).
25. Van Hijum, S. a F. T. *et al.* A generally applicable validation scheme for the assessment of factors involved in reproducibility and quality of DNA-microarray data. *BMC genomics* **6**, 77 (2005).
26. Garcia de la Nava, J., Van Hijum, S. & Trelles, O. PreP: gene expression data pre-processing. *Bioinformatics* **19**, 2328–2329 (2003).
27. Baldi, P. & Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics (Oxford, England)* **17**, 509–19 (2001).
28. Jain, A. K., Murty, M. N. & Flynn, P. J. Data clustering: a review. *ACM Computing Surveys* **31**, 264–323 (1999).
29. Blom, E.-J. *et al.* DISCLOSE : DISsection of CLusters Obtained by SEries of transcriptome data using functional annotations and putative transcription factor binding sites. *BMC bioinformatics* **9**, 535 (2008).
30. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25–9 (2000).
31. Boyle, E. I. *et al.* GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)* **20**, 3710–5 (2004).
32. Blom, E.-J. *et al.* FIVA: Functional Information Viewer and Analyzer extracting biological knowledge from transcriptome data of prokaryotes. *Bioinformatics (Oxford, England)* **23**, 1161–3 (2007).
33. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics (Oxford, England)* **23**, 257–8 (2007).

34. Heijden, V. F. Van Der & Ridder, D. De *Classification, parameter estimation, and state estimation: an engineering ... Journal of Time Series Analysis* **32**, 194–194 (John Wiley and Sons: 2004).
35. Brouwer, R. W. W., Kuipers, O. P. & Van Hijum, S. a F. T. The relative value of operon predictions. *Briefings in bioinformatics* **9**, 367–75 (2008).
36. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297 (1995).
37. Breiman, L. *Random Forests*. *Machine Learning* **45**, 5–32 (Springer Netherlands: 2001).
38. Wolf, Y. I., Rogozin, I. & Kondrashov, A. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res. *Genome Research* **11**, 356–372 (2001).
39. Okuda, S., Katayama, T., Kawashima, S., Goto, S. & Kanehisa, M. ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic acids research* **34**, D358–62 (2006).
40. Price, M. N., Huang, K. H., Arkin, A. P. & Alm, E. J. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome research* **15**, 809–19 (2005).
41. Lawrence, J. G. & Roth, J. R. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843–60 (1996).
42. Romero, P. R. & Karp, P. D. Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics (Oxford, England)* **20**, 709–17 (2004).
43. Siefert, J., Martin, K., Abdi, F., Widger, W. & Fox, G. Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *Journal of molecular evolution* **45**, 467–72 (1997).
44. Ermolaeva, M. D. Prediction of operons in microbial genomes. *Nucleic Acids Research* **29**, 1216–1221 (2001).
45. Overbeek, R. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences* **96**, 2896–2901 (1999).
46. Yan, B., Methé, B. A., Lovley, D. R. & Krushkal, J. Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family *Geobacteraceae*. *Journal of theoretical biology* **230**, 133–44 (2004).
47. Carpentier, A.-S., Riva, A., Tisseur, P., Didier, G. & Hénaut, A. The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA. *Computational biology and chemistry* **28**, 3–10 (2004).
48. Itoh, T., Takemoto, K., Mori, H. & Gojobori, T. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular biology and evolution* **16**, 332–46 (1999).
49. Westover, B. P., Buhler, J. D., Sonnenburg, J. L. & Gordon, J. I. Operon prediction without a training set. *Bioinformatics (Oxford, England)* **21**, 880–8 (2005).

50. Dam, P., Olman, V., Harris, K., Su, Z. & Xu, Y. Operon prediction using both genome-specific and general genomic information. *Nucleic acids research* **35**, 288–98 (2007).
51. Yada, T., Nakao, M., Totoki, Y. & Nakai, K. Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics (Oxford, England)* **15**, 987–93 (1999).
52. Craven, M., Page, D., Shavlik, J., Bockhorst, J. & Glasner, J. A probabilistic learning approach to whole-genome operon prediction. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **8**, 116–27 (2000).
53. Salgado, Moreno-Hagelsieb, G. & Smith, T. Operons in Escherichia coli: Genomic analyses and predictions. *Proceedings of the National Academy of Sciences* **6**, 6652–7 (2000).
54. Moreno-Hagelsieb, G. & Collado-Vides, J. Operon conservation from the point of view of *Escherichia coli*, and inference of functional interdependence of gene products from genome context. *In silico biology* **2**, 87–95 (2002).
55. Moreno-Hagelsieb, G. & Collado-Vides, J. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S329–36 (2002).
56. Sabatti, C., Rohlin, L., Oh, M.-K. & Liao, J. C. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic acids research* **30**, 2886–93 (2002).
57. Tjaden, B. *et al.* Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic acids research* **30**, 3732–8 (2002).
58. Bockhorst, J., Craven, M., Page, D., Shavlik, J. & Glasner, J. A Bayesian network approach to operon prediction. *Bioinformatics (Oxford, England)* **19**, 1227–35 (2003).
59. Chen, X., Su, Z., Xu, Y. & Jiang, T. Computational prediction of operons in *Synechococcus* sp. WH8102. *Genome informatics. International Conference on Genome Informatics* **15**, 211–22 (2004).
60. Chen, X. *et al.* Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome. *Nucleic acids research* **32**, 2147–57 (2004).
61. De Hoon, M. J. L., Imoto, S., Kobayashi, K., Ogasawara, N. & Miyano, S. Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 276–87 (2004).
62. Paredes, C. J., Rigoutsos, I. & Papoutsakis, E. T. Transcriptional organization of the *Clostridium acetobutylicum* genome. *Nucleic acids research* **32**, 1973–81 (2004).
63. Steinhauser, D., Junker, B. H., Luedemann, A., Selbig, J. & Kopka, J. Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics (Oxford, England)* **20**, 1928–39 (2004).

64. Wang, L., Trawick, J. D., Yamamoto, R. & Zamudio, C. Genome-wide operon prediction in *Staphylococcus aureus*. *Nucleic acids research* **32**, 3689–702 (2004).
65. De Hoon, M. J. L., Makita, Y., Nakai, K. & Miyano, S. Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS computational biology* **1**, e25 (2005).
66. Edwards, M. T., Rison, S. C. G., Stoker, N. G. & Wernisch, L. A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context. *Nucleic acids research* **33**, 3253–62 (2005).
67. Jacob, E., Sasikumar, R. & Nair, K. N. R. A fuzzy guided genetic algorithm for operon prediction. *Bioinformatics (Oxford, England)* **21**, 1403–7 (2005).
68. Price, M. N., Huang, K. H., Alm, E. J. & Arkin, A. P. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic acids research* **33**, 880–92 (2005).
69. Janga, S. C., Lamboy, W. F., Huerta, A. M. & Moreno-Hagelsieb, G. The distinctive signatures of promoter regions and operon junctions across prokaryotes. *Nucleic acids research* **34**, 3980–7 (2006).
70. Zhang, G. G., Cao, Z. Z., Luo, Q. Q., Cai, Y. & Li, Y. Operon prediction based on SVM. *Computational biology and chemistry* **30**, 233–40 (2006).
71. Bergman, N. H., Passalacqua, K. D., Hanna, P. C. & Qin, Z. S. Operon prediction for sequenced bacterial genomes without experimental information. *Applied and environmental microbiology* **73**, 846–54 (2007).
72. Charaniya, S. *et al.* Transcriptome dynamics-based operon prediction and verification in *Streptomyces coelicolor*. *Nucleic acids research* **35**, 7222–36 (2007).
73. Tran, T. T. *et al.* Operon prediction in *Pyrococcus furiosus*. *Nucleic acids research* **35**, 11–20 (2007).
74. Tatusov, R. L., Koonin, E. V & Lipman, D. J. A genomic perspective on protein families. *Science (New York, N.Y.)* **278**, 631–7 (1997).
75. Ermolaeva, M. D., Khalak, H. G., White, O., Smith, H. O. & Salzberg, S. L. Prediction of transcription terminators in bacterial genomes. *Journal of molecular biology* **301**, 27–33 (2000).
76. Kingsford, C. L., Ayanbule, K. & Salzberg, S. L. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome biology* **8**, R22 (2007).
77. De Hoon, M. J. L. *et al.* Predicting gene regulation by sigma factors in *Bacillus subtilis* from genome-wide data. *Bioinformatics (Oxford, England)* **20 Suppl 1**, i101–8 (2004).
78. Barrett, T. *et al.* NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic acids research* **35**, D760–5 (2007).
79. Parkinson, H. *et al.* ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic acids research* **35**, D747–50 (2007).

80. Gama-Castro, S. *et al.* RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic acids research* **36**, D120–4 (2008).
81. Taboada, B., Verde, C. & Merino, E. High accuracy operon prediction method based on STRING database scores. *Nucleic acids research* **38**, e130 (2010).
82. Taboada, B., Ciria, R., Martinez-Guerrero, C. E. & Merino, E. ProOpDB: Prokaryotic Operon DataBase. *Nucleic acids research* **40**, D627–31 (2012).
83. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research* **41**, D808–15 (2013).
84. Faith, J. J. *et al.* Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic acids research* **36**, D866–70 (2008).
85. Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic acids research* **29**, 22–8 (2001).
86. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40**, D109–14 (2012).
87. Conant, G. C. & Wolfe, K. H. GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics (Oxford, England)* **24**, 861–2 (2008).
88. Grant, J. R. & Stothard, P. The CGView Server: a comparative genomics tool for circular genomes. *Nucleic acids research* **36**, W181–4 (2008).
89. Kerkhoven, R., Van Enckevort, F. H. J., Boekhorst, J., Molenaar, D. & Siezen, R. J. Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics (Oxford, England)* **20**, 1812–4 (2004).
90. Lulko, A. T., Buist, G., Kok, J. & Kuipers, O. P. Transcriptome analysis of temporal regulation of carbon metabolism by CcpA in *Bacillus subtilis* reveals additional target genes. *Journal of molecular microbiology and biotechnology* **12**, 82–95 (2007).
91. Leenhouts, K. *et al.* A general system for generating unlabelled gene replacements in bacterial chromosomes. *Molecular & general genetics : MGG* **253**, 217–24 (1996).
92. Kuipers, O. P., Beerthuyzen, M. M., Siezen, R. J. & De Vos, W. M. Characterization of the nisin gene cluster *nisABTCIPR* of *Lactococcus lactis*. Requirement of expression of the nisA and nisI genes for development of immunity. *European journal of biochemistry / FEBS* **216**, 281–91 (1993).
93. Kuipers, O. P. *et al.* Transcriptome analysis and related databases of *Lactococcus lactis*. *Antonie van Leeuwenhoek* **82**, 113–22 (2002).
94. Kilstrup, M. Proteomics of *Lactococcus lactis*: phenotypes for a domestic bacterium. *Methods of biochemical analysis* **49**, 149–78 (2006).

95. Neves, A. R., Pool, W. A., Kok, J., Kuipers, O. P. & Santos, H. Overview on sugar metabolism and its control in *Lactococcus lactis* - the input from in vivo NMR. *FEMS microbiology reviews* **29**, 531–54 (2005).
96. Notebaart, R. A., Van Enckevort, F. H. J., Francke, C., Siezen, R. J. & Teusink, B. Accelerating the reconstruction of genome-scale metabolic networks. *BMC bioinformatics* **7**, 296 (2006).
97. Oliveira, A. P., Nielsen, J. & Förster, J. Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC microbiology* **5**, 39 (2005).
98. Voit, E., Neves, A. R. & Santos, H. The intricate side of systems biology. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 9452–7 (2006).
99. Andersen, A. Z. *et al.* The metabolic pH response in *Lactococcus lactis*: an integrative experimental and modelling approach. *Computational biology and chemistry* **33**, 71–83 (2009).
100. Neves, A. R. *et al.* Towards enhanced galactose utilization by *Lactococcus lactis*. *Applied and environmental microbiology* **76**, 7048–60 (2010).
101. Larsen, R., Van Hijum, S. a F. T., Martinussen, J., Kuipers, O. P. & Kok, J. Transcriptome analysis of the *Lactococcus lactis* ArgR and AhrC regulons. *Applied and environmental microbiology* **74**, 4768–71 (2008).
102. Den Hengst, C. D. *et al.* The *Lactococcus lactis* CodY regulon: identification of a conserved cis-regulatory element. *The Journal of biological chemistry* **280**, 34332–42 (2005).
103. Zomer, A. L., Buist, G., Larsen, R., Kok, J. & Kuipers, O. P. Time-resolved determination of the CcpA regulon of *Lactococcus lactis* subsp. *cremoris* MG1363. *Journal of bacteriology* **189**, 1366–81 (2007).
104. Beyer, N. H., Roepstorff, P., Hammer, K. & Kilstrup, M. Proteome analysis of the purine stimulon from *Lactococcus lactis*. *Proteomics* **3**, 786–97 (2003).
105. Kilstrup, M., Jacobsen, S., Hammer, K. & Vogensen, F. K. Induction of heat shock proteins DnaK, GroEL, and GroES by salt stress in *Lactococcus lactis*. *Applied and environmental microbiology* **63**, 1826–37 (1997).
106. Sperandio, B. *et al.* Sulfur Amino Acid Metabolism and Its Control in *Lactococcus lactis* IL1403. *Society* **187**, (2005).
107. Guédon, E., Sperandio, B., Pons, N., Ehrlich, S. D. & Renault, P. Overall control of nitrogen metabolism in *Lactococcus lactis* by CodY, and possible models for CodY regulation in Firmicutes. *Microbiology (Reading, England)* **151**, 3895–909 (2005).
108. Barrière, C. *et al.* Fructose utilization in *Lactococcus lactis* as a model for low-GC gram-positive bacteria: its regulator, signal, and DNA-binding site. *Journal of bacteriology* **187**, 3752–61 (2005).
109. Fallico, V., Ross, R. P., Fitzgerald, G. F. & McAuliffe, O. Genetic response to bacteriophage infection in *Lactococcus lactis* reveals a four-strand approach involving induction of membrane stress proteins, D-alanylation of the cell wall, maintenance of proton motive force, and energy conservation. *Journal of virology* **85**, 12032–42 (2011).

110. Kim, E. B., Piao, D. C., Son, J. S. & Choi, Y. J. Cloning and characterization of a novel tuf promoter from *Lactococcus lactis* subsp. *lactis* IL1403. *Current microbiology* **59**, 425–31 (2009).
111. Kleine, L. L., Monnet, V., Pechoux, C. & Trubuil, A. Role of bacterial peptidase F inferred by statistical analysis and further experimental validation. *HFSP journal* **2**, 29–41 (2008).
112. De Jong, A., Hansen, M. E., Kuipers, O. P., Kilstrup, M. & Kok, J. The Transcriptional and Gene Regulatory Network of *Lactococcus lactis* MG1363 during Growth in Milk. *PloS one* **8**, e53085 (2013).
113. Blom, E.-J., Ridder, A. N. J. a, Lulko, A. T., Roerdink, J. B. T. M. & Kuipers, O. P. Time-Resolved Transcriptomics and Bioinformatic Analyses Reveal Intrinsic Stress Responses during Batch Culture of *Bacillus subtilis*. *PloS one* **6**, e27160 (2011).
114. Ontology, T. C. G. Gene Ontology : tool for the. *Gene Expression* **25**, 25–29 (2000).
115. Fisher, R. A. *Statistical methods for research workers*. 356 (1938).
116. Trip, H., Mulder, N. L. & Lolkema, J. S. Cloning, expression, and functional characterization of secondary amino acid transporters of *Lactococcus lactis*. *Journal of bacteriology* **195**, 340–50 (2013).
117. Shajani, Z., Sykes, M. T. & Williamson, J. R. Assembly of bacterial ribosomes. *Annual review of biochemistry* **80**, 501–26 (2011).
118. Ueta, M. *et al.* Ribosome binding proteins YhbH and YfiA have opposite functions during 100S formation in the stationary phase of *Escherichia coli*. *Genes to cells : devoted to molecular & cellular mechanisms* **10**, 1103–12 (2005).
119. Polikanov, Y. S., Blaha, G. M. & Steitz, T. A. How hibernation factors RMF, HPF, and YfiA turn off protein synthesis. *Science (New York, N.Y.)* **336**, 915–8 (2012).
120. Kilstrup, M., Hammer, K., Ruhdal Jensen, P. & Martinussen, J. Nucleotide metabolism and its control in lactic acid bacteria. *FEMS microbiology reviews* **29**, 555–90 (2005).
121. Van Hijum, S. A. F. T., García de la Nava, J., Trelles, O., Kok, J. & Kuipers, O. P. MicroPreP: a cDNA microarray data pre-processing framework. *Applied bioinformatics* **2**, 241–4 (2003).
122. R Development Core Team, R. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* **1**, 409 (2011).
123. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).
124. Martinussen, J., Sørensen, C., Jendresen, C. B. & Kilstrup, M. Two nucleoside transporters in *Lactococcus lactis* with different substrate specificities. *Microbiology (Reading, England)* **156**, 3148–57 (2010).
125. Dressaire, C. *et al.* Transcriptome and proteome exploration to model translation efficiency and protein stability in *Lactococcus lactis*. *PLoS computational biology* **5**, e1000606 (2009).

126. Redon, E., Loubiere, P. & Cocaign-Bousquet, M. Transcriptome analysis of the progressive adaptation of *Lactococcus lactis* to carbon starvation. *Journal of bacteriology* **187**, 3589–92 (2005).
127. Nouaille, S. *et al.* Transcriptomic response of *Lactococcus lactis* in mixed culture with *Staphylococcus aureus*. *Applied and environmental microbiology* **75**, 4473–82 (2009).
128. Maligoy, M., Mercade, M., Cocaign-Bousquet, M. & Loubiere, P. Transcriptome analysis of Lactococcus lactis in coculture with *Saccharomyces cerevisiae*. *Applied and environmental microbiology* **74**, 485–94 (2008).
129. De Smet, R. & Marchal, K. Advantages and limitations of current network inference methods. *Nature reviews. Microbiology* **8**, 717–29 (2010).
130. Furey, T. S. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics* (2012).doi:10.1038/nrg3306
131. Kaufmann, K. *et al.* Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nature protocols* **5**, 457–72 (2010).
132. Thomas-Chollier, M. *et al.* A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature protocols* **7**, 1551–68 (2012).
133. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–10 (1990).
134. Li, L. GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *Journal of computational biology : a journal of computational molecular cell biology* **16**, 317–29 (2009).
135. Romeo, Y. *et al.* Osmoregulation in *Lactococcus lactis*: BusR, a transcriptional repressor of the glycine betaine uptake system BusA. *Molecular microbiology* **47**, 1135–47 (2003).
136. Pons, P. & Latapy, M. Computing communities in large networks using random walks (2005).
137. Fruchterman, T. M. J. & Reingold, E. M. Graph Drawing by Force-directed Placement. **21**, 1129–1164 (1991).
138. Yamada, T., Letunic, I., Okuda, S., Kanehisa, M. & Bork, P. iPath2.0: interactive pathway explorer. *Nucleic acids research* **39**, W412–5 (2011).
139. Jendresen, C. B., Martinussen, J. & Kilstrup, M. The PurR regulon in *Lactococcus lactis* - transcriptional regulation of the purine nucleotide metabolism and translational machinery. *Microbiology (Reading, England)* **158**, 2026–38 (2012).
140. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **2**, 28–36 (1994).

141. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic acids research* **37**, D26–31 (2009).
142. Geier, F., Timmer, J. & Fleck, C. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC systems biology* **1**, 11 (2007).
143. Gupta, R. *et al.* A computational framework for gene regulatory network inference that combines multiple methods and datasets. *BMC systems biology* **5**, 52 (2011).
144. Sperandio, B. *et al.* Sulfur amino acid metabolism and its control in *Lactococcus lactis* IL1403. *Journal of bacteriology* **187**, 3762–78 (2005).
145. Pinto, J. P. C. *et al.* pSEUDO, a genetic integration standard for *Lactococcus lactis*. *Applied and environmental microbiology* **77**, 6687–90 (2011).
146. Brouwer, R. W. W., Van Hijum, S. A. F. T. & Kuipers, O. P. MINOMICS: visualizing prokaryote transcriptomics and proteomics data in a genomic context. *Bioinformatics (Oxford, England)* **25**, 139–40 (2009).
147. Perkins, T. T. *et al.* A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS genetics* **5**, e1000569 (2009).
148. Van Riel, B. *et al.* A Novel Complex, RUNX1-MYEF2, Represses Hematopoietic Genes in Erythroid Cells. *Molecular and cellular biology* **32**, 3814–22 (2012).
149. Passalacqua, K. D. *et al.* Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis*. *PloS one* **7**, e43350 (2012).
150. Vandernoot, V. A. *et al.* cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. *BioTechniques* **53**, 373–80 (2012).
151. Reddy, J. S. *et al.* Transcriptome profile of a bovine respiratory disease pathogen: *Mannheimia haemolytica* PHL213. *BMC bioinformatics* **13 Suppl 1**, S4 (2012).
152. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**, 139–40 (2010).
153. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562–78 (2012).
154. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106 (2010).
155. Carvalho, R. H. *et al.* Genome-wide DNA methylation profiling of non-small cell lung carcinomas. *Epigenetics & chromatin* **5**, 9 (2012).
156. Taiwo, O. *et al.* Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature protocols* **7**, 617–36 (2012).
157. Veening, J.-W., Smits, W. K. & Kuipers, O. P. Bistability, epigenetics, and bet-hedging in bacteria. *Annual review of microbiology* **62**, 193–210 (2008).

158. Beilharz, K. *et al.* Control of cell division in *Streptococcus pneumoniae* by the conserved Ser/Thr protein kinase StkP. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E905–13 (2012).
159. Van de Werken, H. J. G. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature methods* **9**, 969–72 (2012).